4-8-2020

Research Data Management from the STEM Perspective: Reproducibility, Data Reuse, Data integration

Melissa D. Clarkson
University of Kentucky, mclarkson@uky.edu

Research data management from the STEM perspective

Reproducibility

Data reuse

Data integration

Melissa Clarkson
Assistant Professor
Division of Biomedical Informatics
UK College of Medicine
mclarkson@uky.edu

Assumptions...

Data is structured as tables (rows and columns) or as triples (subject–predicate–object)

Data (and the analysis) is to be shared with your community

Three big ideas in STEM data management:

Reproducibility requires scripted pipelines

Data reuse requires metadata and documentation

Data integration requires use of standards

Datasets have issues.

People producing and using data have issues.

Datasets have issues.

People producing and using data have issues.

Can this data analysis be reproduced?

Can I get the dataset used in this analysis?

Is *this* the dataset used in the analysis?

What is the meaning of the column names in this dataset?

Datasets have issues.

People producing and using data have issues.

Who created this dataset? When?

How do I cite this dataset?

Can I do anything I want with this dataset?

Does this thing in this dataset mean the same as that thing in that other dataset?

Reproducible pipelines for data analysis address some of these issues

- the data
- the analysis
- the software

Reproducible pipelines for data analysis address some of these issues

- the data

- the analysis

- the software

.... and the publication

The goal of the **FAIR data management principles** is for data to become **“first class” assets**

The goal of the **FAIR data management principles** is for data to become **“first class” assets**

F findable

A accessible

I interoperable

R reusable

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES

» Research data

» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*[#]

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem

The FAIR data management principles...

emphasize the importance of **metadata**

should enable both FAIR use by **both humans and machines**

are broadly applicable to **“research objects”**

To be Findable:

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata

F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource

To be Findable:

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata

F3. metadata clearly and explicitly include the identifier of the data it describes

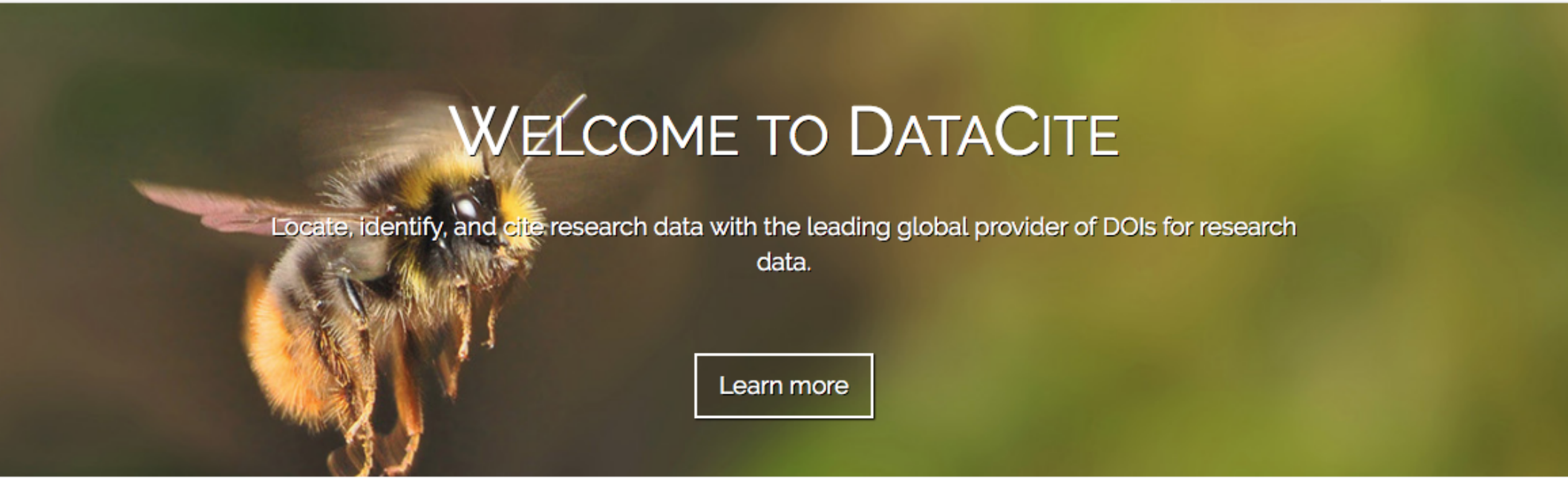
F4. (meta)data are registered or indexed in a searchable resource

DOI Registration Agencies

Current Registration Agencies are listed below, with links to the organizations for requesting more information. New Registration Agencies are being added, and suggestions for additional sector coverage are welcome.

See also a chart showing the areas of coverage provided by each RA, and some examples of services provided by several of the RAs.

Airiti, Inc.	Crossref
China National Knowledge Infrastructure (CNKI)	DataCite



Welcome to DataCite

Locate, identify, and cite research data with the leading global provider of DOIs for research data.

Learn more

Find what you're looking for by searching millions of records with extensive, reliable metadata.

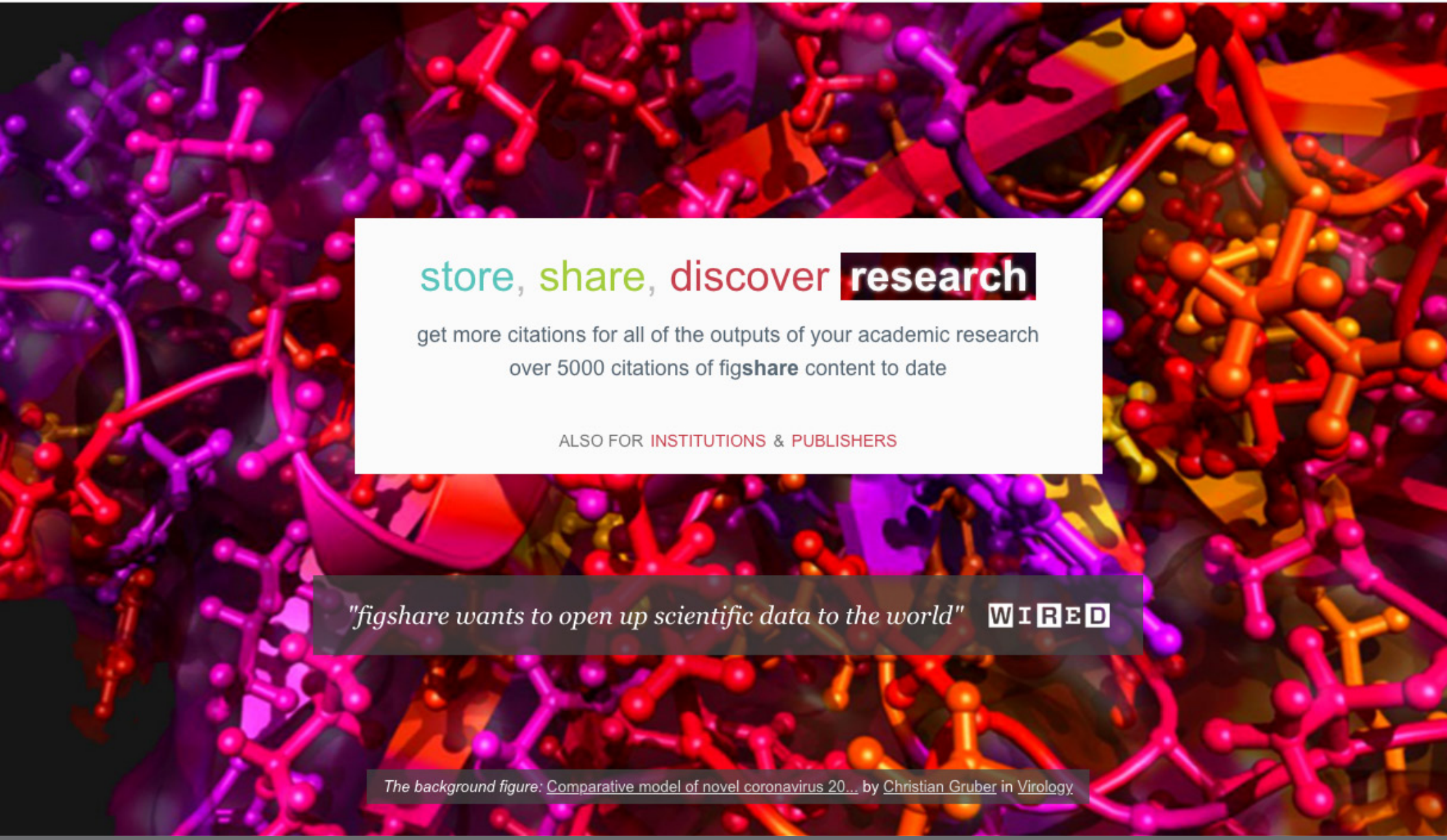
Share your data and reuse the data of others to create the highest impact in the research community.

Cite your research sources with confidence, and receive proper credit when your work is reused.

Connect your research – publications, datasets, software, authors, institutions, and funding data all in one place.

Get started with DataCite!

re3data



figshare.com

Sign in

Dataset Search

Search for Datasets

Try boston education data or weather site:noaa.gov

Learn more about including your datasets in Dataset Search.

datasetsearch.research.google.com

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

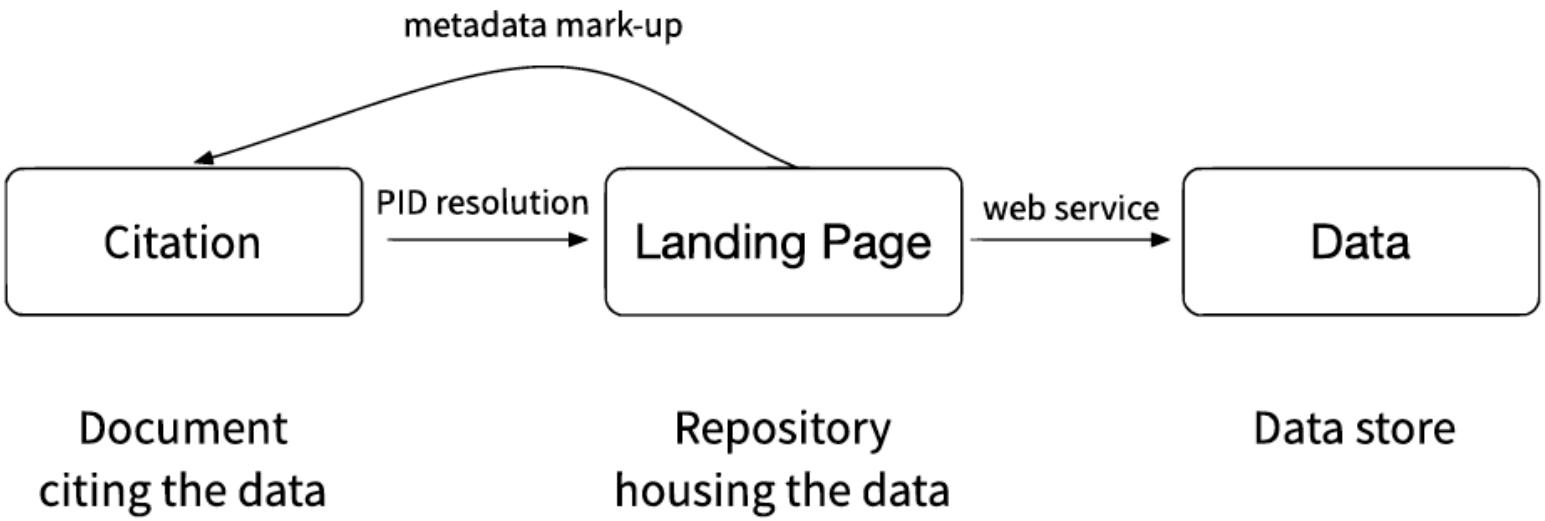


Fig. 1 Generic data citation - relationships of the citation reference, repository landing page and underlying data.

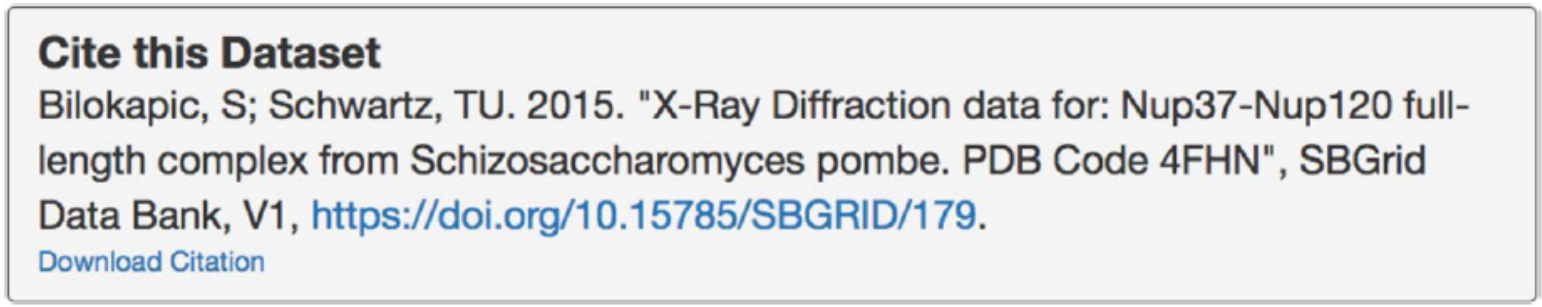
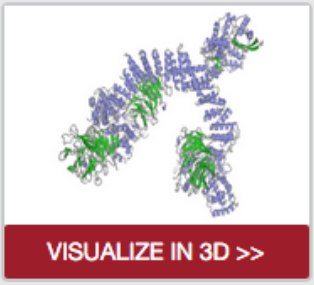


Fig. 2 Providing information about how a dataset should be cited, with download link for citation (in BibTex or other standard bibliographic reference manager format).

"A scholarly citation roadmap for scholarly data repositories"
Scientific Data (2019) Fenner et al. DOI: 10.1038/s41597-019-0031-8

X-Ray Diffraction data from Nup37-Nup120 full-length complex from Schizosaccharomyces pombe, source of 4FHN structure



Data DOI: 10.15785/SBGRID/179 | ID: 179
Publication DOI: 10.1073/pnas.1205151109
4FHN Coordinates: Viewer, PDB (RCSB) (PDBe), MMDB
Schwartz Laboratory, Massachusetts Institute of Technology
Release Date: Oct. 6, 2015

Data Access Instructions

1. If this dataset is locally available, it should be accessable at `/programs/datagrid/179`

2. To download this dataset, please run the following command from your Terminal on a Linux or OS X workstation:

`'rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/179 .'` (Harvard Medical School, USA)

Depending on your location, faster access may be available from a Tier 1 site closer to your location

`'rsync -av rsync://sbgrid.icm.uu.se/10.15785/SBGRID/179 .'` (Uppsala University, Sweden)

`'rsync -av rsync://sbgrid.pasteur.edu.uy/10.15785/SBGRID/179 .'` (Institut Pasteur de Montevideo, Uruguay)

`'rsync -av rsync://sbgrid.ncpss.org/10.15785/SBGRID/179 .'` (Shanghai Institutes for Biological Sciences, China)

3. After the transfer is completed, please issue the following command to verify data integrity:

`'cd 179 ; shasum -c files.sha'`

Storage requirements: 7.1G

Biological Sample:	Nup37-Nup120 full-length complex from Schizosaccharomyces pombe
Dataset Type:	X-Ray Diffraction

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

About Principles Ontologies Resources Citation Participate FAQ Legacy

Search Ontobee Submit

The OBO Foundry

The Open Biological and Biomedical Ontology (OBO) Foundry is a collective of ontology developers that are committed to collaboration and adherence to shared principles. The mission of the OBO Foundry is to develop a family of interoperable ontologies that are both logically well-formed and scientifically accurate. To achieve this, OBO Foundry participants voluntarily adhere to and contribute to the development of an evolving set of principles including open use, collaborative development, non-overlapping and strictly-scoped content, and common syntax and relations, based on ontology models that work well, such as the Gene Ontology (GO).

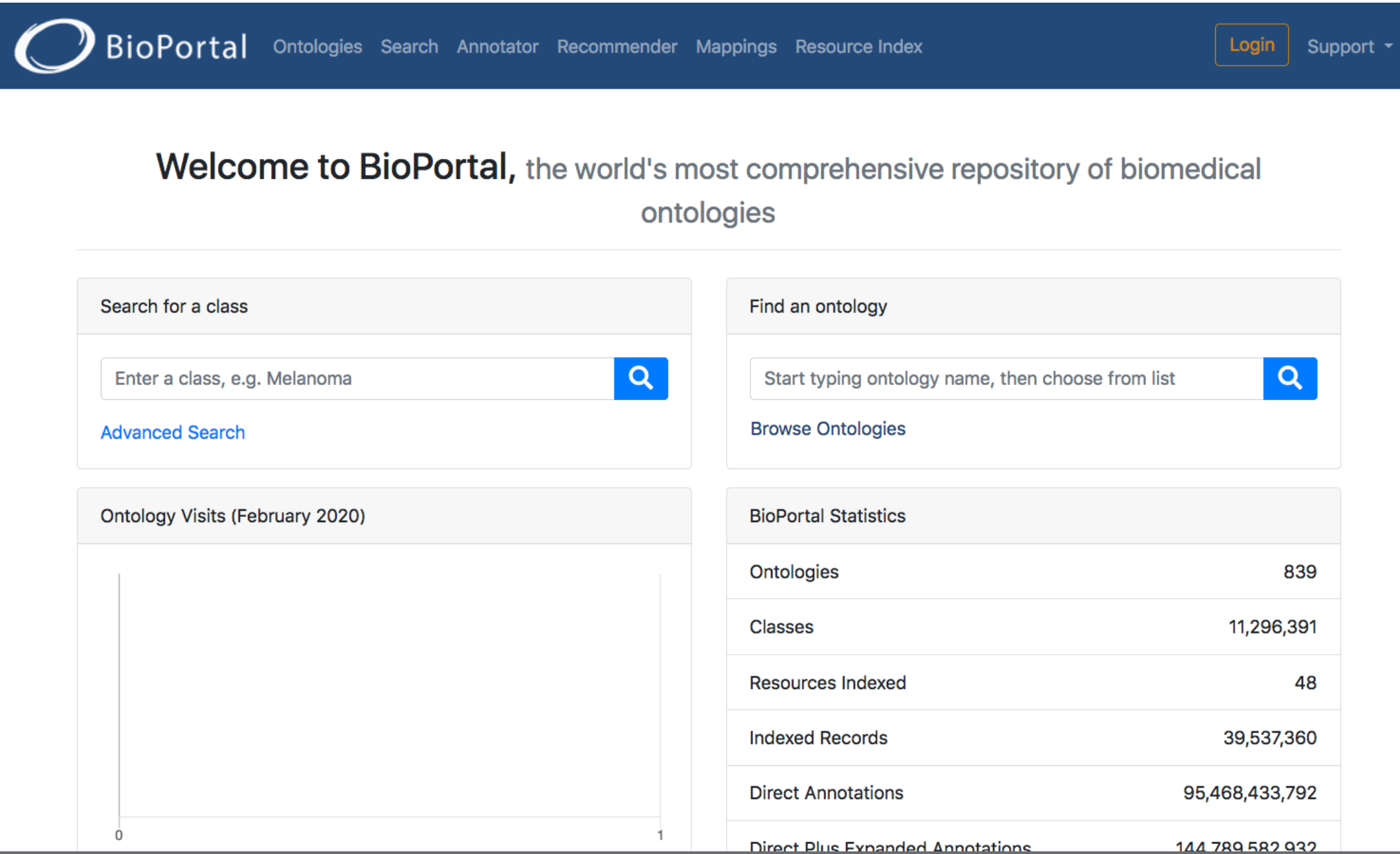
The OBO Foundry is overseen by an Operations Committee with Editorial, Technical and Outreach working groups. The processes of the Editorial working group are modelled on the journal refereeing process. A complete treatment of the OBO Foundry is given in "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration".

On this site you will find a table of ontologies, available in several formats, with details for each, and documentation on OBO Principles.

You can contribute to this site using GitHub OBOFoundry/OBOFoundry.github.io or get in touch with us by joining our mail list https://groups.google.com/forum/#!forum/obo-discuss.

Download table as: [YAML | JSON-LD | RDF/Turtle]

bfo	Basic Formal Ontology (cc) BY	The upper level ontology upon which OBO Foundry ontologies are built. Detail	
chebi	Chemical Entities of Biological Interest (cc) BY	A structured classification of molecular entities of biological interest focusing on 'small' chemical	



bioportal.bioontology.org

FAIRsharing.org
standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

A curated, informative and educational resource on data and metadata *standards*, inter-related to *databases* and data *policies*.

HOW CAN WE HELP?

We guide consumers to discover, select and use these resources with confidence, and producers to make their resource more discoverable, more widely adopted and cited.

Journal editors & publishers

Create and maintain an interrelated list of citable standards, databases and repositories to recommend to your authors, users or their community, and revise this recommendation over time...
[read more]

Researchers | Developers & Curators | Journal Publishers | Librarians & Trainers | Societies & Alliances | Funders

fairsharing.org

To be Reusable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with detailed provenance

R1.3. (meta)data meet domain-relevant community standards

Search

Standards Resources Training Community Publications About

Document, Discover and Interoperate

The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. DDI is a free standard that can document and manage different stages in the research data lifecycle, such as conceptualization, collection, processing, distribution, discovery, and archiving. Documenting data with DDI facilitates understanding, interpretation, and use -- by people, software systems, and computer networks. Use DDI to **D**ocument, **D**iscover, and **I**nteroperate!

Explore Specification

Tools

Learn

Collaborate

Why Use DDI?

- ✔ Generate interactive codebooks
- ✔ Implement data catalogs
- ✔ Build question banks
- ✔ Create concordance mappings
- ✔ Harmonize and compare data
- ✔ Manage longitudinal data sets

Find Out More!

Featured DDI Adopters

What's New

ddialliance.org

W3C Working Group Note

PROV-Overview

An Overview of the PROV Family of Documents

W3C Working Group Note 30 April 2013

This version:
http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
Latest published version:
http://www.w3.org/TR/prov-overview/
Previous version:
http://www.w3.org/TR/2013/WD-prov-overview-20130312/
Editors:
Paul Groth, VU University Amsterdam
Luc Moreau, University of Southampton

Abstract

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used
reliability or trustworthiness. The PROV Family of Documents defines a model, corresponding serializations and other supporting
operable interchange of provenance information in heterogeneous environments such as the Web. This document provides an ov

Status of This Document

w3c.org/TR/prov-overview

Help us build a vibrant, collaborative global commons

Donate Now

Discover the new CC Search

Try the new CC image search with over 300 million images from 19 collections and easier attribution.

Start searching

WHAT'S HAPPENING

Smithsonian Releases 2.8 Million Images + Data into the Public Domain Using CC0

FAIR data management principles

F findable

A accessible

I interoperable

R reusable

Learn more...

“The FAIR Guiding Principles for scientific data management and stewardship”
Scientific Data (2016) Wilkinson et al.
DOI: 10.1038/sdata.2016.18

“A design framework and exemplar metrics for FAIRness”
Scientific Data (2018) Wilkinson et al.
DOI: 10.1038/sdata.2018.118

“A scholarly citation roadmap for scholarly data repositories”
Scientific Data (2019) Fenner et al.
DOI: 10.1038/s41597-019-0031-8